CONTENTS

1. INTRODUCTION 3
OVERVIEW OF STUDY 3
ROUTE GUIDE FOR USERS OF WAVE 1 DATA AND DOCUMENTATION 3
2. STUDY RELATED INFORMATION 4
DESIGN OVERVIEW 4
FIGURE 1. TIMING OF MAINSTAGE AND INNOVATION PANEL (IP) DATA COLLECTION 5
DATA COLLECTION 5
THE PLAYERS, WHO DOES WHAT 5
GETTING READY FOR WAVE 1 5
INTERVIEWERS 6
FIELDWORK 6
PANEL MAINTENANCE 7
RESPONSE OUTCOMES 7
DATA PROCESSING 8
DOCUMENTATION OF THE QUESTIONNAIRES: MODULES AND QUESTIONS 8
CHANGES TO THE QUESTIONNAIRE 9
FIGURE 1. MARK-UP OF HOUSEHOLD QUESTIONNAIRE 10
FIGURE 2. MARK-UP OF INDIVIDUAL LEVEL QUESTIONNAIRE WITH LOOPING 10
OTHER FIELDWORK MATERIALS 11
SAMPLE DESIGN 11
GENERAL POPULATION SAMPLE COMPONENT. 11
GENERAL POPULATION COMPARISON SAMPLE COMPONENT. 12
ETHNIC MINORITY BOOST SAMPLE. 12
SAMPLE STATUS AND FOLLOWING RULES 13
WEIGHTING ADJUSTMENTS IN UK LONGITUDINAL HOUSEHOLD STUDY *UNDERSTANDING SOCIETY* WAVE ONE 14
SELECTING THE CORRECT WEIGHT FOR YOUR ANALYSIS 15
NOT USING WEIGHTS 16
NAMING CONVENTIONS FOR WEIGHTING VARIABLES 17
TECHNICAL DETAILS OF WEIGHTING 17
ENUMERATED INDIVIDUAL WEIGHT 20
IMPUTATION OF INCOME VARIABLES 22
WHAT DO WE IMPUTE? 23
IMPUTATION PROCEDURES 23
ITEM NON-RESPONSE ON INCOME VARIABLES IN THE INDIVIDUAL QUESTIONNAIRE 23
ITEM NON-RESPONSE FOR INCOME VARIABLES IN THE PROXY QUESTIONNAIRE 26

INDIVIDUAL NON-RESPONDENTS WITH NO PROXY QUESTIONNAIRE *26*
COMPUTING TOTAL NET INDIVIDUAL AND HOUSEHOLD INCOME *26*
CODING 27
FILE INFORMATION 27
PRESERVING CONFIDENTIALITY 27
WORKING WITH THE DATA FILES 28
EXAMPLE 1: DISTRIBUTING HOUSEHOLD LEVEL INFORMATION TO THE INDIVIDUAL LEVEL *28*
EXAMPLE 2: SUMMARISING INDIVIDUAL LEVEL INFORMATION AT THE HOUSEHOLD LEVEL *29*
EXAMPLE 3: MATCHING INDIVIDUALS WITHIN A HOUSEHOLD *29*
3. VARIABLE INFORMATION – OVERVIEW BASIC AND DERIVED VARIABLES 31
VARIABLE NAMING AND LABELLING CONVENTIONS 31
LEARNING ABOUT THE STUDY VARIABLES 31
IDENTIFIERS AND USEFUL VARIABLES 31
TABLE 3. SOME USEFUL VARIABLES *32*
DOCUMENTATION OF DERIVED VARIABLES 33
PARADATA IN WAVE 1 33
4. DATA ACCESS 33
CITATIONS AND ACKNOWLEDGEMENTS 34
5. REFERENCES 35

UNDERSTANDING SOCIETY: UK HOUSEHOLD LONGITUDINAL STUDY: WAVE 1, 2009-2010 USER MANUAL

1. INTRODUCTION

OVERVIEW OF STUDY

Understanding Society -- the UK Household Longitudinal Study--is a longitudinal survey of the members of approximately 40,000 households in the United Kingdom (England, Scotland, Wales and Northern Ireland). Households recruited at the first round of data collection are visited one year later to collect information on changes to their household and individual circumstances. Interviews are carried out face-to-face in respondents' homes by trained interviewers. Wave 1 data collection took place between January 2009 and January 2011.

Understanding Society is funded by the Economic and Social Research Council and with funding from multiple government departments (the Department for Work and Pensions, the Department for Education, the Department for Transport, the Department for Culture, Media and Sport, the Department for Communities and Local Government, the Department of Health, the Scottish Government, the Welsh Assembly Government, the Northern Ireland Executive, the Department for Environment, Food and Rural Affairs, and the Food Standards Agency). The scientific leadership team is from the Institute for Social and Economic Research (ISER) of the University of Essex, the University of Warwick, and the Institute of Education, University of London. Professor Nick Buck is the principal investigator. Fieldwork is conducted by the National Centre for Social Research (NatCen) with collaboration with the Central Survey Unit of the Northern Ireland Statistics and Research Agency (NISRA) in Northern Ireland.

The overall purpose of *Understanding Society* is to provide high quality longitudinal data about subjects such as health, work, education, income, family, and social life to help understand the long term effects of social and economic change, as well as policy interventions designed to impact upon the general well-being of the UK population.

ROUTE GUIDE FOR USERS OF WAVE 1 DATA AND DOCUMENTATION

This release has data for the General Population and the Ethnic Minority boost (EMB) sample. Former participants of the British Household Panel Survey (BHPS) are part of *Understanding Society* from Wave 2 (http://www.iser.essex.ac.uk/bhps). The BHPS is a household panel survey of around 8,000 households in the UK, which

has completed 18 annual waves of data collection and has been run by ISER since it began in 1991. Data from the BHPS can be obtained from the UK Data Archive (SN5151 British Household Panel Study, Waves 1-18, 1991-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151.) Data from the Innovation Panel, a separate survey intended to support methodological research (http://www.understandingsociety.org.uk/design/innovation/default.aspx). Data from the Innovation Panel has been released through the UK Data Archive (SN6849 *Understanding Society:* Innovation Panel, Waves 1-2, 2008-2009, http://www.esds.ac.uk/findingData/snDescription.asp?sn=6849).

The Ethnic Minority Boost sample was undertaken to produce enough cases to analyse households and individuals from five major ethnic groups in the UK. The boost sample receives an additional five minutes of questions related to content areas that may particularly involve them. The General Population Comparison sample component is also asked these questions.

As an introduction to the data and documentation we recommend the following reading:

1. How to read the questionnaires, notes on naming conventions, and key variables.

2. The description of the sample design, weighting, and fieldwork procedures and outcomes.

3. Variable level descriptions of the data can be found on the *Understanding Society* website (http://data.understandingsociety.org.uk/documentation/). The online documentation has extensive links between questions and detailed views of variables and datafiles. There is also a search facility for searching questions, variables, modules, and datafiles.

4. The example Stata code for matching variables from different records.

In assembling the documentation, we have drawn upon the documentation for the British Household Panel Survey (Taylor 2010) (see also http://www.iser.essex.ac.uk/bhps).

2. Study Related Information

Design overview

Understanding Society is a panel survey of households with yearly interviews. Data collection for a single wave is scheduled across 24 months. The study begins with a representative probability sample of households. There is an extended discussion of sample design below and in (Lynn, 2009). Adult household members (age 16 or older) are asked questions and the same individuals are re-interviewed in successive years to see how things have changed. There is a short self-completion youth questionnaire for 10-15 year old household members. Children become eligible for a full interview once they reach the age of 16.

The overall study has multiple sample components. In the mainstage survey there is the a) General Population Sample, with its subset the General Population Comparison Sample, b) the Ethnic Minority Boost Sample, and c) participants from the British Household Panel Study. The instruments for the first three components are the same except the EMB sample and the General Population Comparison sample have five additional minutes of questions specifically relevant to the ethnic minority community e.g. ethnic identity and remittances. In addition, there is a separate survey, the Innovation Panel (IP), which is fielded in the year before the mainstage survey. It tests varying measurement issues, and its instruments are somewhat different from the mainstage survey. The timing of data collection for the first two years of the mainstage and IP surveys is shown in Figure 1.

FIGURE 1. TIMING OF MAINSTAGE AND INNOVATION PANEL (IP) DATA COLLECTION

2008				2009				2010				2011				2012			
Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4
IP1																			
				WAVE 1 Mainstage															
				IP2															
								WAVE 2 Mainstage											

Q - Quarter

DATA COLLECTION

THE PLAYERS, WHO DOES WHAT

ISER, together with NatCen and the Central Survey Unit of NISRA, work closely together on all aspects of data collection, implementing an agreed set of survey procedures designed to ensure adequate response and effective data quality. NISRA collaborates with NatCen and is responsible for fieldwork in Northern Ireland.

NatCen manages fieldwork, editing, and coding and data-entry. It also advises on the design of all research instruments. Primary responsibility for design work remains with ISER.

ISER plays a major role in quality control through specification of fieldwork practices; survey materials; editing and coding requirements; and subjecting fieldwork progress to detailed weekly scrutiny. An agreed set of survey-specific procedures to ensure adequate response and effective data quality reinforces this working relationship. Full details of these, and other technical aspects of the data collection and fieldwork, coding, and data processing are found in the *Technical Reports,* published on the *Understanding Society* website (see http://data.understandingsociety.org.uk/.)

GETTING READY FOR WAVE 1

Prior to the first wave of the main *Understanding Society* survey, there were two small pilot studies and a dress rehearsal. A cognitive pilot of 70 individuals was

conducted March – April 2008 to test screening and other core questions, in particular with respect to the ethnicity strand. A translation pilot was conducted in June 2008: 50 interviews were carried out using Bengali and Punjabi translations of the questionnaire to see if there were problems with the operation of the translation program or problems with interviewing with the translated interviews. A survey of 100 households conducted August-September 2008 served as a dress rehearsal of the data collection instruments and procedures.

INTERVIEWERS

For Wave One, 911 interviewers were employed to cover 3,517 areas in the sample. Because of the demanding nature of *Understanding Society* special attempts were made to use interviewers of above average levels of experience and ability. In Northern Ireland, the majority of interviewers had worked on the Northern Ireland component of the BHPS (the Northern Ireland Household Panel Survey), and so were familiar with the design and structure of *Understanding Society*.

In addition to general interviewer training, interviewers working on *Understanding Society* attended a survey-specific face-to-face briefing. These one-day briefings had morning sessions devoted to fieldwork procedures, including dealing with the administrative forms to record contact information, and how to deal with the complexities of multiple dwelling units and multiple households. The afternoon was spent discussing the survey content and reviewing and working with the Blaise computer aided personal interview (CAPI) instrument. Generally around 12-20 interviewers attended each briefing, along with two or three briefing managers or area managers. The briefings were led by at least one researcher from NatCen with the majority also attended by ISER staff. The briefings took place across the UK; Belfast, Birmingham, Brentwood, Bristol, Derby, Edinburgh, Glasgow, Leeds, London and Manchester.

FIELDWORK

The Wave One mainstage fieldwork started on 8th January 2009 and ended on the 7th March 2011 (including the re-issue period).

Before contacting any of their sample, interviewers mailed an introductory card from ISER to all sampled addresses (addressed to "The Occupier"), together with a small leaflet outlining the purpose of the survey. The interviewer called within a week of the mailing. At the end of the first interview, all participating households received a more detailed brochure, giving further information about the survey and thanking respondents for participating.

A minimum of six calls was made at each sampled address before it was considered a non-contact; interviewers were encouraged to make further calls, if possible. A special conversion letter was sent to households which had refused to participate or had not been contacted if there was a potential for success. In total, interviews were achieved in 30,169 households (26,089 in the general population sample, 4,080 in the ethnic minority boost sample), with full or proxy interviews with 50,994 individuals (43,674 in the general population sample and 7,320 in the ethnic minority boost sample).

Interviewers uploaded their work daily, including information about all the calls they have made, whether or not there was any response. This information was collated by NatCen and a weekly field progress monitor report was sent to ISER. Post interview

quality control is carried out with a telephone recall on 10% of all completed interviews.

PANEL MAINTENANCE

ISER maintains a database of information on respondents' location, which builds on contact information collected during the survey interviews, updated throughout the year. This database is the basis on which all fieldwork documents for successive waves of the survey are prepared. As a result of this work, we can better plan in relation to the issuing of the sample for interviews in the next wave. There are, for example, new addresses, household splits and moves out of the country or into an institution.

Prior to fieldwork for Wave 2 mainstage, a summary report of research findings is sent to all adults (except refusals). This mailing also has an address-confirmation slip. The letter for the Wave 1 inter-wave report included a unique invitation code to allow the sample member to register with the Participants website. The participant website can be seen at http://participants.understandingsociety.org.uk/.

Change of address cards were also returned to ISER in cases where a whole household moved or a new resident returned the card giving the forwarding address. Finally, it is possible for ISER to be notified of some deaths through this means.

RESPONSE OUTCOMES

The tables below present the household and individual response rates for Wave 1. The individual response rates are for co-operating households only.

Table 1. Household response rates among eligible households

	General Population Sample			Ethnic Minority Boost
	Great Britain	Northern Ireland	Total	
Productive	57.4%	61.7%	57.6%	52.0%
Non-contact	4.1%	5.1%	4.2%	25.4%
Refusal	36.5%	32.7%	36.3%	35.5%
Other	2.1%	0.5%	2.0%	7.1%
N	43232	2093	45325	10111

The response rates for the ethnicity boost sample component makes a correction for the probability of non-interviewed cases being ineligible.

Table 2. Individual response rates: Wave 1

	General Population Sample			Ethnic Minority Boost
	Great Britain	Northern Ireland	Total	
Full interview	82.0%	77.3%	81.8%	72.4%
Proxy interview	5.3%	3.5%	5.2%	6.9%
Refusal	6.5%	9.2%	6.7%	8.7%
Other non-interview	6.1%	9.9%	6.3%	12.1%
n	47615	2584	50199	9237

DATA PROCESSING

Data from each sample month are delivered by NatCen to ISER in batches. The delivery is scheduled for 4 months following the beginning of the fieldwork process, to allow time for interview re-issue, coding, and data entry from paper documents, e.g. the self-completion instruments.

Data is delivered as SPSS system files, which are then exported to triple-S data exchange format and imported into a SIR database.

Quality control processes include extensive data checking to ensure that the data conform to the expected structure and the routing and range constraints defined by the questionnaire specifications. Data distributions are also checked for theoretical and statistical plausibility. This checking is done through direct scrutiny and by analyses which 'road-test' the data.

Data anomalies are investigated to determine whether they are related to:
1) the invalid specification of the questionnaire; 2) the incorrect scripting of the questionnaire; 3) a failure to specify that a particular constraint should be included in the questionnaire; 4) an incorrect implementation of the check, or; 5) a problem in exporting and/or delivering the data. After investigation, steps may include correcting the specification, data editing, reporting the error to NatCen to be fixed in a subsequent delivery and/or a quality feedback report suggesting changes to the questionnaire or field practice in subsequent waves.

Batch-specific databases are merged into a single database, from which anonymised data is exported for the creation of public use files.

DOCUMENTATION OF THE QUESTIONNAIRES: MODULES AND QUESTIONS

The text of the questionnaires in pdf format is part of the documentation provided through the UK Data Archive. Questionnaires can also be found at: http://data.understandingsociety.org.uk/documentation/mainstage/questionnaires/ . The documentation is for the mainstage survey (household and individual) and the adult and youth self-completion instruments. The instruments are an important source of information about the wording of individual questions, who was asked, and what questions precede and follow.

Most of the interview is conducted with a computer-assisted personal interview (CAPI). The CAPI instrument governs the flow of questions and recording of answers, but it is not convenient for documentation. On the study website, we are presenting the questionnaire in different formats, which have different advantages and disadvantages. For example, the PDF versions are useful for printing sections of the instrument. The self-completion instruments are shown in two formats, pdf to correspond to the way they appeared to participants and in a format annotated with variable names. In addition, the Address Record Forms may be seen with fieldwork materials on the website: http://data.understandingsociety.org.uk/documentation/mainstage/fieldwork-documents/.

The questionnaires are organised in modules. Modules can be searched for in the online documentation system. In the pdf-formatted questionnaire, clicking on entries in the table of contents will advance you to the beginning of that module. In addition, the questionnaire can be searched for variable names or any word of interest.

Instruments and survey materials were translated into multiple languages: Bengali; Punjabi in Urdu and Gurmukhi scripts; Welsh; Arabic; Somali; Cantonese; Urdu and Gujarati. Translated documents can be requested by email from info@understandingsociety.org.uk.

CHANGES TO THE QUESTIONNAIRE

Questionnaire changes have been made under certain circumstances. At the end of the first six months of data collection in Wave 1, multiple variables were dropped because of the length of the interview, e.g. cutting of the employment history module. At the same time other modifications were made, e.g. in question format. Notes about these changes can be seen in the online documentation system in the variable view.

Figure 1 shows a marked up sample page, providing information for how to interpret the questionnaire text. Note that the variable names in the questionnaire do not have the wave prefix (a_).

FIGURE 1. MARK-UP OF HOUSEHOLD QUESTIONNAIRE

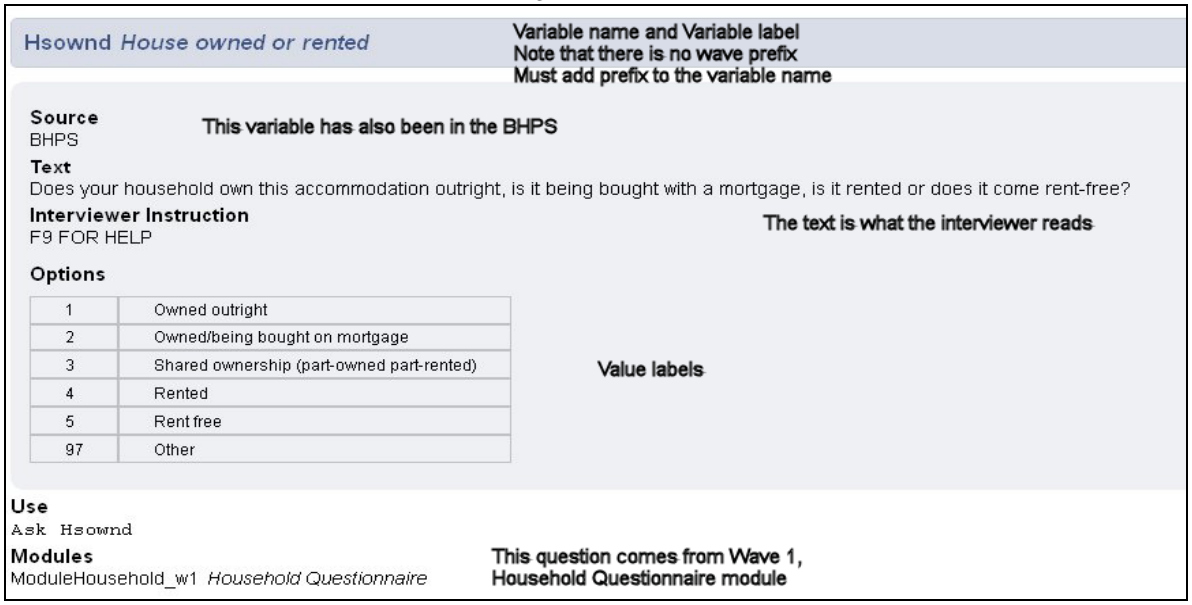
Hsownd House owned or rented
Variable name and Variable label
Note that there is no wave prefix
Must add prefix to the variable name
Source
BHPS
This variable has also been in the BHPS
Text
Does your household own this accommodation outright, is it being bought with a mortgage, is it rented or does it come rent-free?
Interviewer Instruction
F9 FOR HELP
The text is what the interviewer reads
Options
1 Owned outright
2 Owned/being bought on mortgage
3 Shared ownership (part-owned part-rented)
4 Rented
5 Rent free
97 Other
Value labels
Use
Ask Hsownd
Modules
ModuleHousehold_w1 Household Questionnaire
This question comes from Wave 1,
Household Questionnaire module

Figure 2 shows a marked up sample page from the individual interview. The question is more complex. The question is asked about each natural or biological child, so multiple variables are associated with the question for each natural child. The variable is located in the datafile a_natchild, which has one record for each natural child.

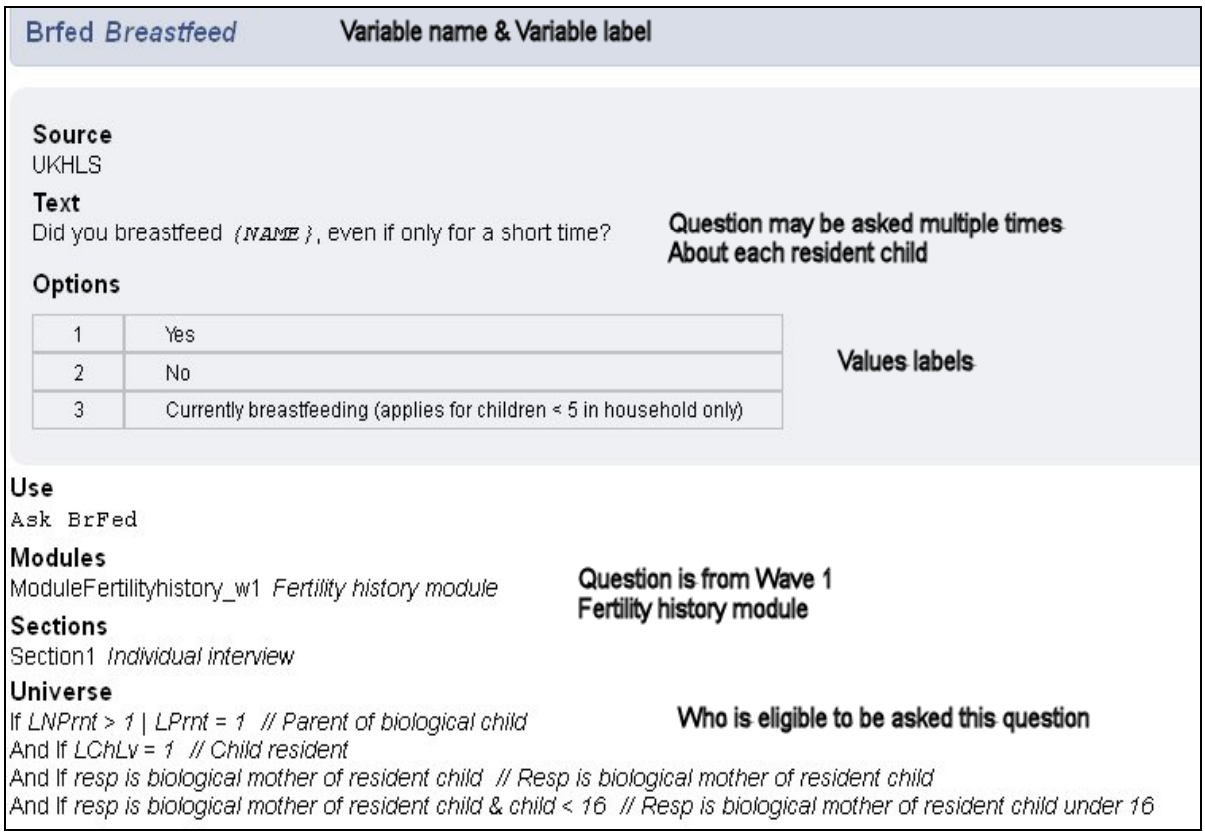
Brfed Breastfeed
Variable name & Variable label
Source
UKHLS
Text
Did you breastfeed (NAME), even if only for a short time?
Question may be asked multiple times
About each resident child
Options
1 Yes
2 No
3 Currently breastfeeding (applies for children < 5 in household only)
Values labels
Use
Ask BrFed
Modules
ModuleFertilityhistory_w1 Fertility history module
Question is from Wave 1
Fertility history module
Sections
Section1 Individual interview
Universe
If LNPrnt > 1 | LPrnt = 1 // Parent of biological child
Who is eligible to be asked this question
And If LChLv = 1 // Child resident
And If resp is biological mother of resident child // Resp is biological mother of resident child
And If resp is biological mother of resident child & child < 16 // Resp is biological mother of resident child under 16

FIGURE 2. MARK-UP OF INDIVIDUAL LEVEL QUESTIONNAIRE WITH LOOPING

OTHER FIELDWORK MATERIALS

Other fieldwork materials are also on the website: http://data.understandingsociety.org.uk/documentation/mainstage/fieldwork-documents/. One example is the Showcards, which are used to help respondents with their answers. Showcards are referenced in the questionnaire. Project Instructions were prepared for interviewer training and to serve as a resource in data collection. Documents for communicating with participants are also included on the website.

The Address Record Form (ARF) is an important source of information about responding and non-responding households. It has the call record, observations on characteristics of accommodation and households, and household outcomes. In Wave 1 there are several different versions of the ARF. The first distinction is between the General Population Sample (GP) and the Ethnic Minority Boost Sample (EB). The versions labelled ARF are longer because they include questions for screening household members for eligibility. ARF's labelled 2 or 3 are for addresses with multiple households and/or dwelling units. Finally, there are versions for ARF EB1 Year 1 or Year 2. This change in form was required by the change in selection criteria implemented in Year 2 of Wave 1 (see Berthoud et al., 2009 for more detail). The ARF screening card was a show card used during the screening interviews. Additional information about completion of the ARF can be found in the Project Instructions for Interviewers. (http://data.understandingsociety.org.uk/documentation/mainstage/fieldwork-docments).

SAMPLE DESIGN

The *Understanding Society* sample consists of a new large general population sample plus four other components (the ethnic minority boost sample, the general population comparison sample, the ex-BHPS sample and the innovation panel sample). The design of all five components is described in more detail in an *Understanding Society* working paper (see Lynn 2009). The general population sample is based upon two separate samples of residential addresses, one for England, Scotland and Wales and one for Northern Ireland. The England, Scotland and Wales sample is a proportionately stratified (equal probability), clustered sample of addresses selected from the Postcode Address File. The Northern Ireland sample is an unclustered systematic random sample of addresses selected from the Land and Property Services Agency list of domestic addresses.

GENERAL POPULATION SAMPLE COMPONENT. The sample for England, Scotland and Wales was selected in two stages. The first stage was to select a sample of postcode sectors to serve as primary sampling units. The second stage was to select addresses within each sampled sector. Prior to selection, any postcode sector with fewer than 500 residential addresses was first grouped with an adjacent sector and thereafter treated as a single sector. The list of all sectors was then sorted into twelve geographical strata, consisting of ten regions in England plus Scotland and Wales as separate strata. Within each of the twelve strata, sectors were sorted into three sub-strata based upon the proportion of household reference persons classified as non-manual workers, based on 2001 Census data. Within each of the 36 sub-strata, sectors were then sorted into three further sub-divisions based on population density (households per hectare) and within each of the 108 resultant

sub-divisions, sectors were listed in order of ethnic minority density. From the sorted list, a systematic random sample of 2,640 sectors was selected, with probability proportional to the number of residential addresses in the sector. These sectors were then allocated systematically to 24 monthly samples, with 110 sectors in each monthly sample. Within each postal sector, 18 addresses were selected using systematic random sampling. The England, Scotland and Wales sample in this data release is therefore based upon an initial sample of 47,520 addresses.

In Northern Ireland, 2,395 addresses were selected in a single stage from the list of domestic addresses. In combination, this data release is therefore based upon a total of 49,915 addresses.

At each address, the final stage of sampling was carried out by field interviewers. This consisted of identifying persons to be defined as sample members. All persons resident at each sample address at the time the interviewer made contact were deemed to be a sample member, with the exception of the small proportion of addresses that contained more than three dwellings or households. In those cases, three dwellings or households were sub-sampled at random.

GENERAL POPULATION COMPARISON SAMPLE COMPONENT. The General Population Comparison Sample (GPCS) has one sampled address for 40% of the selected postal sectors in General Population Sample (GPS) component for Great Britain. In other words, of the 2,640 general population sectors, 60% of them (1,584) contain 18 GPS addresses and the other 40% contain 17 GPS addresses and one GPCS address. The persons in these households will be designated as members of the General Population Comparison sample, regardless of ethnic group membership. Members of the General Population Comparison sample are a random subsample of the General Population Sample component and they should be included in analyses of the General Population Sample component.

ETHNIC MINORITY BOOST SAMPLE. The Ethnic Minority Boost Sample was designed to provide at least 1,000 adults from each of five groups: Indian, Pakistani, Bangladeshi, Caribbean, and African.

The initial step was identifying postal sectors with relatively high proportions of relevant ethnic minority groups, based upon 2001 Census data and more recent Annual Population Survey data. The set of 3,145 sectors constituted approximately 35% of the sectors in Great Britain and covered between 82% and 93% of the population of the five ethnic minority groups.

The 3,145 sectors were sorted into four strata based on the expected number of ethnic minority households that would be identified by the sampling and screening procedures (see Berthoud et al., 2009 for details). All sectors were included for the stratum where a yield of three or more households was expected. In the other three strata, sectors were sub-sampled at rates of 1 in 4, 1 in 8, or 1 in 16 respectively. This was done to constrain the number of sectors that might have just one or two eligible sample households (or even none). The total number of postal sectors selected for inclusion in the ethnic minority boost sample was 771. Of these 6 were in Scotland, 7 were in Wales, and the remaining 758 were in England, with a concentration in London (412 sectors).

The number of addresses selected per postal sector ranged from 15 to 103. Sampling fractions varied across the sectors in a way designed to deliver target numbers of respondents in each target ethnic minority group with adequate statistical efficiency (see Berthoud et al., 2009 for more details). In sectors selected for both the General Population Sample component and the Ethnic Minority Boost sample, a single systematic sample of the required total number of addresses was selected and allocated in a systematic way to the two sample components, thus ensuring that both sample components are spread throughout the whole sector.

The final stage of sampling was done by the interviewers for the Ethnic Minority Boost sample, though its procedures were somewhat more complex. You can see the steps described in the Project Instructions for Interviewers (http://data.understandingsociety.org.uk/assets/476). At addresses containing more than three dwellings or households, the procedures to sub-select dwellings or households were as described above for the General Population Sample component. Within each household, rather than all resident persons becoming sample members, there were three additional steps:

- A "screen" was carried out to identify whether there were any persons from target ethnic groups in the household.
- A random mechanism was applied to certain target groups identified by the screen in order to select only a desired proportion into the sample (non-mixed Indian, African, Far Eastern, Middle Eastern). For other target groups, all resident persons were included in the sample (mixed Indian, Bangladeshi, mixed Caribbean, Sri Lankan, Chinese, Turkish).
- In households included in the sample in the previous two steps, all members of target ethnic groups were deemed to be members of the Ethnic Minority Boost sample (including children). All persons of other ethnic groups are not Ethnic Minority Boost sample members. They will be interviewed as temporary sample members for so long as they remain co-resident with at least one Ethnic Minority Boost sample member.

The overall sampling fractions combine a) the probability of sampling the sector, b) the fraction of addresses selected within the sector, and c) the probability of a household being retained following the application of the random selection mechanism described above.

SAMPLE STATUS AND FOLLOWING RULES

There are three possible sample statuses: Original Sample Members (OSMs), Temporary Sample Members (TSMs), and Permanent Sample members (PSMs). The definitions are as follows.

Original Sample Members (OSMs).

All members of *Understanding Society* Innovation Panel and General Population Sample households enumerated at wave 1, including absent household members and those living in institutions who would otherwise be resident, are Original Sample Members (OSMs). All ethnic minority members of an enumerated household eligible for inclusion in the Ethnic Minority Boost sample are OSMs. Any child born to an OSM mother after wave 1 and observed to be co-resident with the mother at the survey wave following the child's birth will be an OSM. OSMs, of all ages, are

followed for interview and remain eligible as long as they are resident within the UK. They remain potentially eligible sample members for the life of survey.

The case may arise where the only OSM in the household is a child. Other household members are then TSMs so long as they are co-resident with the child, and therefore eligible for interview, even if the child is not yet old enough to be eligible for interview. If the OSM child moves house, they are followed to their new address and those living with the OSM child are eligible for interview. If the OSM child moves into an institution, where normally just the OSM/PSM would be interviewed and not co-residents, a split-off household is created containing only the OSM child and the household enumeration grid completed. The child OSM is an eligible sample member, even if they are not eligible for interview because of their age.

Temporary Sample Members (TSMs).

Any members of an enumerated household eligible for inclusion in the Ethnic Minority Boost sample at wave 1 who are not from a qualifying ethnic minority are Temporary Sample Members (TSMs) at wave 1. This was the only category of TSM at wave 1. Any new person found to be co-resident in an OSM or PSM household after wave 1 is a TSM. This would include any child born to an OSM father after wave 1 but not an OSM mother and observed to be co-resident with the father at the survey wave following the child's birth. TSMs remain eligible for interview as long as co-resident in an OSM/PSM household. TSMs who are not co-resident in an OSM/PSM household are not followed and become ineligible for interview. TSMs are identified as re-joiners if they are subsequently found in an OSM/PSM household and then become eligible for interview.

Permanent Sample Members (PSMs).

PSMs are TSMs who are followed for interview after they no longer live with an OSM. This is done for substantive research reasons because of the additional contextual information they may provide for the analysis of OSMs. At present, there is only one category of PSM, but others may be defined in the future. Any TSM father of an OSM child born after wave 1 and observed to be co-resident with the child at the survey wave following the child's birth is a PSM. PSMs remain potentially eligible for interview for the life of survey.

WEIGHTING ADJUSTMENTS IN UK LONGITUDINAL HOUSEHOLD STUDY *UNDERSTANDING SOCIETY* WAVE ONE

A number of weights are provided for data users in order to adjust for unequal selection probabilities, nonresponse and potential sampling error. Importantly, household-level weighted analysis will correctly take into account the boost in Northern Ireland, the Ethnic Minority (EM) boost, and will adjust for household level nonresponse. For individual-level analysis, in addition to the above adjustments, weighted analysis will adjust for within-household nonresponse at wave 1 and will match (poststratify) the sample to population estimates on sex, age and geographical region variables (GOR). Considering the complexity of the study design, weights should be selected carefully following advice provided below.

WARNING:

- NEVER use GPS or EMBOOST part of the sample separately. If you aim to study the general population, use all available data with the weight suitable for your analysis. The weights are designed for the two parts of the sample to be used in combination for substantive research of any subgroup of the population (including analysis of only ethnic minority groups or that which aims to represent the whole population).
- NEVER conduct unweighted analysis if you aim to generalise your results to the UK population. For advanced users who want to model nonresponse on their own, we provide design weights (see below) which adjust the sample for unequal selection probability. Note, that adjusting for first wave nonresponse is different from adjusting for attrition and requires variables which have values for both responding households and never responding households.

SELECTING THE CORRECT WEIGHT FOR YOUR ANALYSIS

Given the complexity and multi-purpose of the study design a number of different weights are provided to meet different needs of users. The weight for your analysis reflects the survey instrument which is the source of the data being used in the analysis, and the analysis level (household or individual).

Analysis level	***Data source***	**Analysis Weight**	**Design Weight (for advanced users only)**
household level	household grid and household interview	a_hhdenus_xw	a_hhdenus_xd
individual level	household grid and household interview	a_psnenus_xw	a_psnenus_xd
individual level	proxy and full interview, 16+	a_indpxus_xw	a_psnenus_xd
individual level	full interview only (no proxy), 16+	a_indinus_xw	a_psnenus_xd
individual level	extra 5 minutes (full interview, 16+)	a_ind5mus_xw	a_ind5mus_xd
individual level	self-completion interview, 16+	a_indscus_xw	a_psnenus_xd
individual level	self-completion youth interview, 10-15	a_ythscus_xw	a_psnenus_xd

Note, that all the weights follow a naming convention which is designed to help users to pick the correct weight. The name of each weight reflects the wave for which the weight is calculated, level of analysis, data source and its nature (design weight, cross-sectional analysis weight or longitudinal analysis weight). The rules are described in the ‘Naming Conventions for Weighting Variables’ section below.

For individual level analysis of adults it is possible that a researcher may want to combine information from different questionnaire sources. In this situation please select the weight suitable for the lowest level according to the hierarchy below:

Level of Analysis	Questions available for
4	household level (all enumerated individuals)
3	proxy and full interview, 16+
2	full interview only (no proxy), 16+
1	self-completion interviews adult or youth

For example, if in one model you use questions available for proxy and full interview as well as for self-completion interview, then the correct weight will be a_indscus_xw – the weight for self-completion interview as its level (1) is lower than the level for proxy and full interview (3).

NOT USING WEIGHTS

Note, that an unweighted analysis does not reflect population estimates correctly unless all the assumptions below are true. It is suggested that researchers publishing or presenting unweighted estimates make these assumptions explicit.

If no weighting is used, your analysis assumes:

1. that all estimates of interest are the same in Northern Ireland as in the rest of the UK;
2. that all estimates of interest are the same among ethnic minority (EM persons) as among nonethnic minority members (non-EM persons);
3. that non-EM persons who live with EM persons in the same household are the same as non-EM persons who don't live with EM persons with respect to your estimates of interest;
4. that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
5. that households who didn't respond in wave 1 are the same with respect to your estimates as households who did respond;
6. for individual-level analysis, that individuals who responded are the same with regard to your estimates as those who didn't respond either at household level or at within-household level.

Using design weights only will correct your model for points 1-4, but not for points 5 and 6 above unless appropriate nonresponse correction is implemented. Note, that nonresponse in *Understanding Society* is more complex than in most of other surveys as many of households selected for EMBOOST are of unknown eligibility.

We, therefore strongly suggest using weighted analyses at all times when analysing data from *Understanding Society,* the UKHLS.

NAMING CONVENTIONS FOR WEIGHTING VARIABLES

Naming conventions have been adopted for weighting variables. This will help users to establish the name of the weight they need or to identify the nature of a weight. The structure is as follows:

w_xxxyyzz_aa

where
w=wave
xxx=target population
yy=instrument
zzz=sample
aa=weight type

xxx:
hhd: household
psn: persons 0+
ind: persons 16+
yth: persons 10-15

yy:
en: enumeration (grid)
in: interview
px: interview or proxy
5m: “extra 5 minutes” items
sc: self-completion
ns: nurse visit
bd: blood

zz:
us: UKHLS sample (GPS and ethnic boost)
bh: BHPS sample
ip: Innovation panel

aa:
lw: longitudinal analysis weight
xw: x-sectional analysis weight
ld: longitudinal design weight
xd: x-sectional design weight

Example

a_indinus_xw is the cross-sectional analysis weight for individual interview data from wave 1, representing the population of persons aged 16+

TECHNICAL DETAILS OF WEIGHTING

Household level weights consist of two components: design weight and nonresponse adjustment for household level nonresponse. Individual level weights consist of four components design weight, nonresponse adjustment for household level nonresponse, individual level within-household nonresponse, and post-stratification to population characteristics. Each of the components is explained below.

Design weight

The design weight corrects for unequal probability of selection at a number of levels.

Household level design weight corrects for:

- Unequal selection probability due to the boost in Northern Ireland. GPS selection probabilities in Northern Ireland are approximately twice those in other parts of the UK;
- Unequal selection probability due to the ethnic minority boost. Selection probabilities in the EMBOOST part of the sample vary considerably between areas, depending on the estimated ethnic mix of the area. Additionally, households in high density areas with at least one ethnic minority member were weighted to account for combined probability to be selected as part of GPS or as part of EMBOOST samples;
- The selection probability of households in a dwelling with more than 3 households or at an address with more than three dwellings is adjusted for the fact that only three such households were selected from the same address.

Individual level design weights correct for all the above with one specific difference: non-EM persons who live with EM persons in the same household have a chance to be selected only via the GPS part of the sample, and not via EMBOOST. This means that non-EM persons in the EMBOOST (who are TSMs) are given a design weight of 0 while non-EM persons in the GPS are given the household design weight. The weights for EM persons adjust for their dual probability to be part of GPS or EMBOOST.

Individual level design weights for the extra five minutes is similar to the above design weight and differs in the following ways. It adjusts for the fact that GPS-comparison sample is only 1/45th of GPS original sample, that all EM members in low-density areas were administered the extra-five minutes, and that EM members in high-density areas had a chance to be selected into either the GPS-comparison sample or the EMBOOST. Similar to the above weight, non-EM persons were assumed to have a chance to be part of only GPS-comparison sample.

Household level Nonresponse Adjustment

Household level nonresponse adjustment is more complex than in other surveys given the large number of households which were selected as part of EMBOOST with unknown eligibility. Households who were selected as part of EMBOOST sample were screened on whether they contain at least one member of a relevant EM group (Berthoud et al., 2009). Given the low proportion of eligible households in EMBOOST sample it is unrealistic to assume that all nonresponding households would be eligible, i.e. contain at least one EM member. To take this into account we modelled eligibility and used this information in household nonresponse adjustments such that household which were more likely to be eligible had a higher influence on nonresponse correction. Note, that predicted eligibility multiplied by design weight is released for all the EMBOOST sample households of unknown eligibility as part of a_hhdenus_xd. This will enable an advanced user to model first wave household nonresponse taking into account chance to be eligible among households of unknown eligibility.

To model eligibility we used predictors from the sampling frame and administrative neighbourhood data linked at a geographical level (for detailed description see below). After excluding ineligible addresses (like businesses or demolished and nonexistent addresses), the eligibility was modelled using only EMBOOST households with known eligibility status (either screened out or screened in). This prediction was then extrapolated onto EMBOOST households of unknown eligibility (e.g. not contacted). Given the limited number of selected addresses in Wales and Scotland and differences between countries in the available auxiliary variables (see below), we predicted eligibility using two models. The first included common predictors for England and Wales and the eligibility was predicted for these two countries. The second was based on England, Wales and Scotland, using a more limited number of predictors and the eligibility was predicted for Scotland only from this model.

Following this, a probability to respond was estimated using backward stepwise logistic regression weighted by eligibility status (where ineligible were excluded, those known to be eligible had a eligibility of one, and those with unknown eligibility had a weight proportional to the predicted probability of being eligible obtained from the above model). The predictors used in this model were the same as for the eligibility model and are described in detail below. Given that administrative neighbourhood data differs between England, Wales, Scotland and Northern Ireland, a separate model was implemented for each country. GPS and EMBOOST response propensity was modelled together (which allowed us to model nonresponse within each country separately), but the indicator of EMBOOST was retained in the model even if not significant.

Predictors used for eligibility model and household level nonresponse correction come from the following sources:

- Sampling frame information, including such variables as sample month and geographical region;
- Predicted ethnic density of postcode sector for 5 main ethnic groups in England, Scotland and Wales (as described in Berthoud et al., 2009);
- A wide range of indicators from Census 2001 and the most updated version of neighbourhood statistics as of summer 2011, linked separately for England, Wales, Scotland and Northern Ireland (see below).

The household nonresponse correction weight was calculated as the inverse of probability from the above model. This weight was multiplied by household design weight to create wave 1 household level weight. Design effect was estimated using this weight showing that no truncation was necessary. The obtained weight was scaled to a mean of 1 and was named a_hhdenus_xw.

Neighbourhood statistics

For England and Wales the information was linked at Middle Layer Super Output Area (MSOA) or Lower Layer Super Output Area (LSOA) level and was obtained from http://neighbourhood.statistics.gov.uk. The examples of linked information obtained from Census 2001 include proportions in MSOA of employed, retired, outright property owners, travellers to work using different types of transport, single household members, households with one car, peoples with different types of

qualification and professional occupation, among others. Other linked information include 2010 information on multiple deprivation indexes, on crime instances, 2009 information on inflow and net change of neighbourhood population, and proportion of different allowance claimants, 2008 information on hospital admissions and energy consumption.

For Scotland the information was linked at datazone level from http://www.scrol.gov.uk/scrol/common/home.jsp and from http://www.scotland.gov.uk/Topics/Statistics/SIMD. From Census 2001 information was obtained on population density, mean age, average household size and number of rooms per household in datazone, as well as proportions in datazone born in Scotland and outside EU, of different religious denomination, employed, unemployed and retired, disabled, those with different levels of qualification and types of occupation, different types of accommodation among others.

For Northern Ireland the information was linked at Super Output Area (SOA) level and was obtained from http://www.ninis.nisra.gov.uk/. Examples of predictors obtained from Census 2001 at SOA level include average hours worked by residents, average age of residents, percentages of residents with different level of qualifications, with different employment statuses, with different types of marital status among others. The predictors also include 2007-2009 information on multiple deprivation indexes.

ENUMERATED INDIVIDUAL WEIGHT

The weight for analysis of enumerated individuals (a_psnenus.xw) is not equivalent to the household weight for all household members, as often happens in other household studies. This is because we have TSMs in wave 1, who are nonethnic members selected into EMBOOST part of the sample. Thus, the individual level design weight is not equal to the household level design weight for individuals in households containing a mix of EM and non-EM persons. The weight for analysis of enumerated individuals is calculated as the product of individual level design weight a_psnenus_xd and household level nonresponse correction (described above). The design effect was then tested showing that no truncation was necessary. Weighted sample distributions were then compared to ONS mid-year estimates (with a correction for institutionalised population) and poststratification was implemented for the full matrix of gender by geographical region by 5-10 year age groups. Thus the individual level enumerated weight consists of:

Individual level design weight * household nonresponse correction * poststratification adjustment

The obtained weight is then scaled to have a mean of one.

Individual Level Nonresponse Adjustment

Five different individual level weights were prepared for users reflecting different levels of nonresponse and different questionnaire instruments. Each individual level weight consists of:

Individual level design weight * household nonresponse correction * individual level nonresponse correction conditional on household response * poststratification adjustment

The individual nonresponse correction (conditional on household nonresponse) is modelled at three levels:

- For age16+ respondents who either completed full main interview or for whom proxy interview was completed (for a_indpxus_xw);
- For age 16+ respondents who completed full main interview only (for a_indinus_xw and a_ind5mus_xw);
- For age 10+ respondents who completed and returned self-completion questionnaire (for a_indscus_xw and a_ythscus_xw).

Note, that the same model was used for respondents regardless of whether they were selected into GPS or EMBOOST; that response propensity is assumed to not depend on whether respondents received extra five-minutes or not; and that conditional on age (present in the model), response to self-completion is assumed to have the same predictors for adults and youth (this assumption allowed modelling response in each country separately, which wouldn't be possible for youth sample otherwise).

The individual level response conditional on household response was modelled using backward stepwise logistic regression separately for England, Wales, Scotland and Northern Ireland. The four models were implemented for each of the three levels described above. The predictors used in the models include all the predictors used for household level nonresponse models and predictors plus both individual and household-level variables obtained from the household questionnaire, such as age and gender, marital and employment status, household size and presence of children in the household, as well as household expenditure on food and food outside, consideration of use of environmental energy, among others.

The individual-level non-response adjustment was obtained as the inverse of the predicted probability and was then multiplied by the relevant (either individual or extra-five minutes) design weight and by the household nonresponse correction. No truncation was deemed necessary as there were no extreme values substantially impacting design effects. The poststratification was implemented as described above in the individual level enumeration weight section, except that a greatly reduced matrix was used in the case of the extra-five-minutes weight, due to the much smaller sample size to which this weight applies. After multiplying by the poststratification adjustment, each of the following five obtained weights was then scaled to a mean of one:

- Age 16+ main and proxy interview respondents (a_indpxus_xw);
- Age 16+ main interview respondents (a_indinus_xw);
- Age 16+ extra-five minutes respondents (a_ind5mus_xw);
- Age 16+ self-completion respondents (a_indscus_xw);
- Age 10-15 self-completion respondents (a_ythscus_xw).

IMPUTATION OF INCOME VARIABLES

Understanding Society collects detailed information on personal income. All individuals aged 16 or more are asked to report

- wages,
- self-employment earnings,
- second job earnings,
- interests and dividends,
- pensions (National Insurance/state retirement pension, pension from a previous employer, pension from a spouse's previous employer, private pension/annuity, widow's or war widow's pension, widowed mother's allowance or widowed pension),
- benefits (severe disablement allowance, disability living allowance, war disablement pension, attendance allowance, carer's allowance, incapacity benefit, income support, job seeker's allowance, national insurance credits, child benefit, child tax credit, working tax credit, maternity allowance, housing benefit, council tax benefit, foster allowance/guardian allowance/rent rebate, rate rebate, employment and support allowance, respond to work credit, sickness and accident insurance, in-work credit for lone parents and pension credit) and
- other income sources (educational grant, trade union and friendly society payment, maintenance or alimony, payments from a family member not living together, amount for rent from boarders or lodgers, rent from any other property).

These personal income variables can be summed to obtain the total personal income. Total household income can be computed from the personal total incomes of all household members.

The difficulty is that some of the income components can be missing. More precisely there can be three types of missing cases:

1. item non-response when individuals respond to the individual questionnaire but do not answer to some or all the questions on income components;
2. individual non-response when individuals fail to respond to the individual questionnaire;
3. household non-response when there is neither a household nor the individual questionnaire response.

We have 59,466 individuals for whom at least the household questionnaire is available, and among these individuals 80.3% provide a personal interview, 5.5% have a proxy interview, whereas 14.2% have neither a proxy nor a personal interview. The item non-response rate for individuals who provide an individual questionnaire varies across income variables. It goes from a maximum of about 50% for self-employment earnings to zero for some of the benefit variables, and it is generally below 20% for the remaining income variables.

WHAT DO WE IMPUTE?

In *Understanding Society* we do not impute income variables for non-responding households. Responding households are households for which the household questionnaire and information on the household composition/structure (household grid module) are available. We suggest that the user take account of household non-response via weighted estimates (described in Weighting Adjustments)

For individuals who respond to the individual questionnaire but do not provide answers to all income questions (item non-response), we impute the following personal income variables: wages, self-employment earnings, second job earnings, interests and dividends, pensions, benefits and others income sources.

For individuals for whom a proxy questionnaire is available, we impute total earnings and total income whenever missing. The proxy questionnaire is a short version of the individual questionnaire with questions on total earnings and total income as well as other variables.

Finally, for individuals in responding households for whom neither the personal nor the proxy questionnaire is available, we impute only the total personal income.

Based on these imputations we can compute total personal and household income for all individuals belonging to responding households.

IMPUTATION PROCEDURES

The procedure used in *Understanding Society* is imputation by chained equations. Each income variable is imputed by stochastic regression imputation using as predictors a large set of auxiliary variables which includes income variables and other potential correlates such as personal and household socio-demographic characteristics. Some of these characteristics are missing and must also be imputed, but the released data contains imputed values only for the income variables.

Imputation by chained equations (ICE) allows for interdependence between income and auxiliary variables by considering univariate models estimated separately and sequentially (see Van Buuren et al 1999 and Ragunathan et al 2001). This method has been already used in some major household panel surveys such as the European Community Household Panel Survey.

In the following we describe the imputation by chained equations (ICE) adopted for item non-response in the individual personal and proxy questionnaires, and for individual non-response, that is, for those for whom there is neither an individual nor a proxy questionnaire available.

ITEM NON-RESPONSE ON INCOME VARIABLES IN THE INDIVIDUAL QUESTIONNAIRE

The imputation of income variables in the individual questionnaire is performed considering a separate equation for each of the income components, except for pensions. Pensions are imputed at aggregate level as total amount of all pensions received.

We use log linear models for each of our income variables. The explanatory variables are a set of characteristics collected in the personal or household questionnaires. The specification of the models varies by income variable but it generally includes the following variables:

- personal socio-economic variables (age, sex, self-reported ethnic group, indicator for respondent born in the UK, marital status, education level, general health, current subjective financial situation);
- personal income variables (excluding the one used a dependent variable);
- household characteristics (number of children in the household, house tenure, house type, household size);
- job characteristics (log number of hours normally worked per week, log number of hours per months in a second job, log years of job tenure, permanent or temporary job, occupation (soc 2000, 1 digit), number employed at the current job workplace (for employees), number of employees if self-employed, whether is self-employed and hires employees, whether the employment organization is private or not (only for employees), type of ownership if self-employed (sole ownership or partnership), an indicator for whether annual business accounts are prepared for the Inland Revenue for tax purposes if self-employed);
- household variables reflecting economic situation (log amount spent on food from food shops in four weeks prior to interview, log amount spent on food eaten outside the home in four weeks prior to interview, log last year expenditure on domestic fuel (e.g. electricity and gas), number of bedrooms in the house, number of other bedrooms in the house, Council Tax band);
- government office regions.

Furthermore, we use additional regression models to impute explanatory variables when missing. More specifically we use log linear regression for continuous variables and binary, ordered and multinomial logit models respectively for dummy, ordinal and unordered categorical variables. Finally, we consider interval regression when we have brackets rather than point information or when we have *a priori* information which allows us to bound the missing income variable. This is the case for: (a) dividends and interests for which we have bracketed information; (b) gross (net) wages and self-employment earnings because we use the corresponding net (gross) income variable as lower (upper) bound; (c) pensions because we impute the total amount of pensions which is given by the sum of different pensions and, in cases where one or more of the pensions are missing, we use the sum of reported pensions as lower bound for the total pension.

The imputation by chained equations, ICE, starts by considering the following recursive (triangular) system of imputation equations,

where $Y_1,Y_2,\ldots,Y_k$ are the income and auxiliary variables to be imputed ordered from the one with the fewest percentage of missing values, Y_1, to the one with the largest percentage of missing values Y_k, X is a set of auxiliary variables observed for all individuals, α's and β's are parameters and $u_1,u_2,\ldots,u_k$ are random errors. Such recursive system allows us to carry out the imputation separately for each variable and sequentially. The sequential procedure is given by the following steps:

1. estimation of the first equation and imputation of the missing values for Y_1,
2. estimation of the second equation using the imputed values to replace the missing values of Y_1, and imputation of Y_2,
3. repetition of estimation and imputation steps sequentially for each of the following equations until when all k variables, $Y_1,Y_2,\ldots,Y_k$ have been imputed.

We use stochastic imputation, that is, we draw the imputed values from the posterior predictive distribution of the variable to be imputed conditional to the observed data. For more details about stochastic imputation we refer to Rubin (1987), Schafer (1997), and Kenward and Carpenter (2007).

This sequential estimation is consistent only if the recursive system is valid. Since this is not necessarily a valid assumption, ICE uses the imputed values produced using the above recursive system as starting values in an iterative imputation process. In other words, the starting values are used to begin a new cycle of imputations where each equation is estimated sequentially, but this time using as explanatory variables both X and all the imputed variables $Y_1,Y_2,\ldots,Y_k$ excluding the one used as dependent variable. At the end of this new cycle, a set of new imputed variables is produced and used to begin a further new cycle of imputations. These cycles of imputations are repeated until convergence. Notice that in practice some of the variables will be imputed by excluding some of the Xs and Ys variables because it does not always make sense to use all variables as predictors.

We split the income variables collected in the individual questionnaire in two subgroups. The first group consists of gross wages, gross self-employment earnings, gross second job earnings, interests and dividends. The second group is net wages, net self-employment earnings, total pensions, and each of the benefits and remaining income sources. Our ICE is performed in two blocks. We start by imputing the first group of variables (gross) by specifying a first block of chained equations. Then we use the imputed values together with the observed ones to perform a second ICE for the second group of income variables (net). Running two ICE sequentially produces consistent results under the assumption that the system of equations for all income variables and explanatory variables can be written as a two-block recursive system. We assume that this is the case because the variables imputed in the second block (mainly net income variables and benefits) are theoretically an almost "deterministic" function of the variables imputed in the first step, while the income variables imputed in the first step (mainly gross earnings variables) should be a function of job and personal characteristics, which we use as predictors. We use stochastic imputation for all variables except the gross self-employment income which is imputed using a deterministic imputation. This choice was because of the huge number of missing cases for self-employment income, which led to a big variance for the residual error.

ITEM NON-RESPONSE FOR INCOME VARIABLES IN THE PROXY QUESTIONNAIRE
The only income variables reported in the proxy questionnaires are the total gross earning and total gross income. We impute missing values for these two variables again using ICE. The imputation is based on the sample of persons responding to the individual questionnaire (where missing values have been replaced with the imputed values produced by ICE as explained in last section) together with the sample of individuals for whom a proxy questionnaire is available. The imputation process is comparable to the one described in last section. Since individuals answering the proxy questionnaires are asked to report income brackets rather than point values, we use interval regressions for both earning and income. We first impute total gross earnings and total gross income using the explanatory variables described above. Then we use the imputed explanatory and gross income variables to impute total net earnings and total net income.

INDIVIDUAL NON-RESPONDENTS WITH NO PROXY QUESTIONNAIRE
For individual non-respondents with no proxy questionnaire but in responding households, we use information from the household questionnaire to impute a total personal income. The procedure used is again the imputation by chained equations (ICE). We first impute the total gross income, then we impute the total net income using gross income as a predictor in addition to the other explanatory variables. The user should notice that the imputation of personal income for individuals for whom there is neither a personal nor a proxy questionnaire is based only on variables available in the household questionnaire. More precisely we use:

- individual socio-economic variables (age, sex, marital status, ethnicity, work)
- household socio-economic variables (household size, number of children in the household, whether there is nobody in the household who speaks English, whether the interview had to be translated, house type, an indicator for whether the person is owner of the house, the external condition of the address relative to the others, number of bedrooms in the house, number of other rooms in the house, value of the property for home owners, number of cars, number of durables, log last year's expenditure on domestic fuel (e.g. electricity and gas), amount spent on food eaten outside the home in four weeks prior to interview, amount spent on food from food shops in four weeks prior to interview, weekly rent paid, whether the household can keep the accommodation warm enough)
- government office region, indicator for whether the area is a low density area for ethnic minorities.

COMPUTING TOTAL NET INDIVIDUAL AND HOUSEHOLD INCOME
Once all personal income variables are imputed we computed the total monthly individual net income. We define all income variables in monthly amounts before this step. Furthermore, since some of the income variables are collected as gross, we had to consider a gross/net conversion.

It seems plausible to assume that interests and dividends and total amount of pensions are reported net. Furthermore, most of the benefits are non-taxable or reported as net. For this reason, we need a conversion from gross to net only the following income items: second job gross earnings, carer's allowance, incapacity

benefit, job seeker allowance, rent from any property (excluding rent from boarders or lodgers), and employment and support allowance. Because we impute both gross and net wages and self-employment earnings, we can use the ratio between net and gross earnings to derive the net amounts from the corresponding gross amounts. We impose a conversion rate of one for individuals with no earnings.

After these conversions, we computed the personal total net income for all individuals who responded to an individual questionnaire. For people who did not respond to the individual questionnaire we compute an imputed total net income (see previous sections). Finally, by adding reported or imputed total net income for all members belonging to the same household we compute the total household net income.

CODING

Occupational coding for respondent's occupations and parental occupations was carried out using the Computer Assisted Standard Occupational Classification (CASOC) system developed by Peter Elias. As a result of the six-figure codes attached via CASOC, matching of the 1990 SOC coding with previous occupational classifications is now possible; in addition, special algorithms within CASOC allow the re-coding of SOC codes into Socio-economic Group (SEG), RGSC (Registrar Generals Social Class), Goldthorpe, Hope-Goldthorpe, Cambridge Scale and ILOISCO 88.

Several questions, e.g. country of birth, religion, political party, national identity, and citizenship had an "other, please specify" option. These responses were coded using an iterative automated process.

Coding was also done for an open-ended question: "We've asked you a lot of questions but we also want to know what has happened in your **own** life that has been **especially important to you**. Can you please tell me anything that has happened to you, or your family, over the past year that has stood out as important? The respondent could give up to four answers. The answers were recorded verbatim and coded for type of event and its subject.

FILE INFORMATION

The data release consists of multiple files in SPSS or Stata formats distributed by the UK Data Archive Economic and Social Data Service. The list of files and their descriptors can be seen in the online documentation system.

PRESERVING CONFIDENTIALITY

In preparing the data for the release we have taken steps to maintain the confidentiality of responses. These include not releasing the full date of birth and not releasing detailed geographic identifiers. Open or narrative text, e.g. names of schools or employers, has not been released since it may indirectly identify individuals.

A Special Licence version of the data will be released through the UK Data Archive. The study has a Data Access Committee to take decisions on applications

requesting access to electronic data and biological samples from *Understanding Society*. Its aim is to allow important research to proceed while minimising risks, particularly to Study participants.

WORKING WITH THE DATA FILES

Understanding Society has data files at the household and individual levels across multiple waves accruing over time. The bulk of the data is at the household or individual level. Even within a single wave there are households, enumerated household members, files with records about each child, spells for the duration of marital or cohabiting partnerships, receipts from specific income sources, and pairs of individuals linked by type of relationship. In future releases of the data, individuals will be followed through time. Some individuals will move from a household and join with other people in a new household. As additional waves of data are released, analysts will then wish to merge data across waves. Whether working cross-sectionally or longitudinally, researchers often want to restructure the data into different levels or units of analysis.

Some planning is important. First, what is the level of the resulting analysis file? Do you want to be working at the household or individual level? Second, what are the identification variables can you use to link the files? Third, what is the type of link or merge? Finally, do you need to make changes to the files you are planning to link or merge? This might include subsetting the sets of variables or renaming variables so the set of variables in the files are distinct.

Example code for matching files

The examples are illustrated with code for Stata and SPSS. The three examples include:

- Distributing household level information to the individual level
- Summarising individual level information at the household level
- Matching individuals within a household

EXAMPLE 1: DISTRIBUTING HOUSEHOLD LEVEL INFORMATION TO THE INDIVIDUAL LEVEL

In this example we will distribute household level information to individuals in those households. We can do this by merging household level file (such as w_hhresp) with an individual level file (such as w_indresp) within the same wave.

Stata code

```
version 11
use pidp a_hidp  a_marstat using a_indresp, clear
** merge in hhsize from HHRESP and save new dataset
merge m:1 a_hidp using a_hhresp,keep(1 3) keepus(a_hhsize) nogen
save final1, replace
```

SPSS code

```
COMMENT open household file and keep household var and identifier.
GET FILE= a_hhresp.sav
    /keep a_hidp a_hsownd.
```

```
SAVE OUTFILE = hhtemp.sav.
SORT CASES by  a_hidp.
COMMENT open individual file, keep selected variables.
GET FILE=a_indresp.sav
/KEEP a_hidp pidp a_mvever.
MATCH FILES file = *
/TABLE = hhtemp.sav  /by a_hidp.
SAVE OUTFILE = a_indmove.sav.
```

EXAMPLE 2: SUMMARISING INDIVIDUAL LEVEL INFORMATION AT THE HOUSEHOLD LEVEL

In this example we will summarise individual level information within a household (number of males in the household) and then match that onto the household level file.

Stata code

```
** get hidp dvage and sex from INDALL for male sex only
use pidp a_hidp a_dvage a_sex using a_indall if a_sex==1, clear
** collapse by hidp summing nmales
ge nmales=1
collapse (sum) nmales,by(a_hidp)
** merge in hhsize from HHRESP and save new dataset
merge m:1 a_hidp using a_hhresp,keep(1 3) keepus(a_hhsize) nogen
save final2, replace
```

SPSS code

```
COMMENT make dummy variable for males from the file indall file.
GET FILE = 'C:\Data\a_indall.sav'
  KEEP = a_hidp pidp a_sex.
COMPUTE maledum = 0.
IF a_sex=1  maledum=1.
SORT CASES by a_hidp.
COMMENT aggregate individual response to household level.
AGGREGATE OUTFILE = * /  BREAK = a_hidp
  /nmales = sum(maledum).
COMMENT match aggregate level file to subset of household responders file.
MATCH FILES file = * / file = 'C:\Data\a_hhresp.sav' / by a_hidp
  / KEEP = a_hidp a_hhsize a_hsownd nmales.
SAVE OUTFILE = 'C:\Data\smallhh.sav'.
```

EXAMPLE 3: MATCHING INDIVIDUALS WITHIN A HOUSEHOLD

In this example we will match the information of wives onto that of their partners/spouses.

Stata code

```
** In this example we will match the information of wives onto that of
their partners/spouses.
** Open the dataset with information on all persons in responding
households and keep only those persons who have a spouse/partner in the
household
use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall if a_hgpart>0, clear
```

```
** rename the prefix a_ to something that would indicate that this
information relates to the spouse or partner
renpfix a_ sp_
** rename the spouse/partner pno variable to the respondent pno variable as
this will be used to match on to the respondent information. Then sort and
save the data
rename sp_hgpart a_pno
rename sp_hidp a_hidp
drop sp_pno
tempfile spousepartner
save `spousepartner'
** again open the data with information on all persons in responding
households
use a_hidp a_pno a_hgpart a_sex a_dvage using a_indall if a_hgpart>0, clear
** rename the prefix a_ to something that would indicate that this
information relates to the respondent
renpfix a_ r_
** as we want to match on a_hidp and a_pno rename r_hidp and r_pno back to
these
rename r_hidp a_hidp
rename r_pno a_pno
** sort and merge with the spouse partner file and save new dataset
sort a_hidp a_pno
merge 1:1 a_hidp a_pno using `spousepartner',nogen
save final3, replace
```

SPSS code

```
COMMENT From indresp select if a spouse or partner and make a matching
variable.
GET FILE = "C:\Data\a_indresp.sav"
/KEEP = a_hidp a_pno pidp  a_sex a_dvage a_jbstat a_hgpart.

SELECT IF (a_hgpart gt 0).
COMPUTE a_matno = a_pno.
SORT CASES by a_hidp a_matno.
COMMENT save as a spouse file and rename to show that they are spouse
variables.
 SAVE OUTFILE = 'C:\Data\spouse.sav'
   /DROP=a_pno a_hgpart
/RENAME = ( pidp a_sex a_dvage a_jbstat = s_pidp s_sex s_age s_jbstat).

COMMENT now the other spouse and match.
GET FILE = "C:\Data\a_indresp.sav"
/KEEP = a_hidp a_pno pidp  a_sex a_dvage a_jbstat a_hgpart.

COMMENT select if there is a spouse or partner and make a matching
variable.
SELECT IF (a_hgpart gt 0).
COMPUTE a_matno = a_hgpart.
SORT CASES by a_hidp a_matno.

COMMENT match files.
MATCH FILES file = */in=regular
   /file = 'C:\Data\spouse.sav'/ in=spouse
  /by a_hidp a_matno.
```

3. VARIABLE INFORMATION – OVERVIEW BASIC AND DERIVED VARIABLES

VARIABLE NAMING AND LABELLING CONVENTIONS

Most variables have a mnemonic name. Variables begin with a prefix designating the wave of data collection ("a_" for the first wave, b_ for the second wave). We have used "W_" to denote waves in general. We have attempted to keep the names of variables that came from the BHPS the same for the convenience of analysts.

Many derived variables are shown by the suffix '_dv'. Derived variables include variables copied over from one file to another for analytic convenience, variables that categorise a particular variable, e.g. age category; variables that combine information from multiple variables, e.g. body mass index from self reported height and weight. Information about how the derived variable is produced is shown in the notes for derived variables in the detailed variable view of the online documentation.

Additional codes denote different types of reasons for the lack of a valid response. These values have not been specified as missing in Stata or SPSS. However, these statistical packages have commands to assign values to missing for many variables simultaneously. Codes are

-9 Missing by error.

-8 Not applicable to the person or because of routing

-7 Proxy respondent. The question was not asked of proxy respondents or derived variable cannot be computed for proxy respondents.

-2 Refused

-1 Don't know

The meaning of other values is explained with the variable's value labels. There may also be Notes in the detailed variable view on the website.

LEARNING ABOUT THE STUDY VARIABLES

There are multiple resources for learning about the study variables in order to plan analyses. These include the questionnaires and the module and variable views in the online documentation system.

Many of the basic (non-derived) variables can be learned about directly from the questionnaires. As was shown in Figure 2, the questionnaire has much useful information. Please note that in the questionnaire, the variable name does not have the wave prefix. It also shows the brief variable label, text of the question, source of the question and value labels. Showcards to help the respondent in answering are also marked as part of the questionnaire. You can go back and forth from the question view to the variable view.

IDENTIFIERS AND USEFUL VARIABLES

Households are identified by w_hidp, a wave specific variable with a different prefix for each wave. It can be used to link information about a household from different

records within a wave, but cannot be used to link information across waves. Sine the composition of households change between waves, the data do not include a longitudinal household identifier.

Individuals are identified by the personal identifier (pidp), which is consistent in all waves and can be used to link information about a person from different records belonging to one wave, or to link information from different waves. Individuals are also identified by w_pno – the person number within the household. The combination of w_hidp and w_pno is unique for each individual.

TABLE 3. SOME USEFUL VARIABLES

Variable	Description
a_hidp	Household identifier
a_hhsize	Household size
a_hsownd, a_tenure_dv	House owned or rented
a_hhtype_dv	Household type
a_fihhmngrs_dv	Gross household income in past 30 days
a_fihhmnnet_dv	Net household income in past 30 days
a_emboost	Ethnic minority boost flag
a_gpcomp	General population sample comparison with EM boost
a_hhresp_dv	Household response outcome
a_hhdenus_xw	Household cross-sectional weight
a_psu_dv	Primary sampling unit
a_strata_dv	Sampling strata
a_hhdenus_xd	Household design weight
pidp	Cross-wave person identifier
a_country	Country or part of the UK
a_gor_dv	Government office region
a_urindew_dv, a_urindsc_dv a_urindni_dv	Urban-rural indicator for England/Wales, Scotland, and Northern Ireland respectively
a_ivfio_dv	individual response outcome
a_jbnssec8_dv	Social class (NS-SEC)
a_sex	Sex
a_dvage	Age
a_marstat	Legal marital status
mpid, fpid	cross wave identifier of natural mother, father
a_nchild_dv	number of natural children in household
a_ethnic1 to a_ethnic14, a_ethnic96	Specific ethnicities plus none of these
a_jbstat	Current economic activity (employment status)
a_jbhas	Did paid work last week
a_ukborn	Born in the UK and UK country of birth
a_fenow	Still in further education
a_qfhigh	Highest educational qualification
a_sf1	General or self-rated health
a_health	Long standing illness or impairment
a_jbsoc00	Current occupation (SOC2000)

a_indpxus_xw	proxy adult interview cross-sectional weight
a_indinus_xw	Interviewed individual cross-sectional weight
a_indscus_xw	Int with self comp individual cross-sectional weight
a_ind5mus_xw	Int + extra 5 min individual cross-sectional weight
a_ythscus_xw	Self-completion youth interview, 10-15

DOCUMENTATION OF DERIVED VARIABLES

Derived variables are variables that are copied from one file to another for analytic convenience, or computed from one or more variables. Some are computed by the Blaise CAPI program to control the routing within the questionnaire. Others were computed for the purpose of analysts. Analysts should consult the description of derived variables that they plan to use in their analyses.

The derived variables are documented on the detailed variable view on the *Understanding* Society website. The documentation summarises the variables used in the computation of the derived variable. See the detailed view for a_scghq2_dv , a categorical or caseness expression of scores for the GHQ-12 as an example.

PARADATA IN WAVE 1

Some paradata, additional data collected about the interview process is available. These consist of call records, timings data and other information collected by the interviewers during the interview. The w_callrec datafile has information on the number of calls made as well as the issue number, time and date and the outcome of each call. This is available in the dataset w_callrec. Information on the date of receipt of the case and the interviewer associated with each issue as well as the outcome at the end of each issue period is available in the file w_issue. Timings are in the w_indresp datafile. Timing variables give the start time for a module. The duration can be calculated in relation to the next module in the questionnaire. Timings are given for the household questionnaires and for individual modules. In addition to this, information collected in the address response form (ARF) by interviewers while contacting each household and requesting household members to participate in the survey is available in w_hhsamp. This includes data on the area surrounding the address, the type of accommodation and other information that the interviewer can observe about sampled addresses. Reasons for refusal are also available. Interviewers also collect some information about the quality of the interview and persons present during the interview process. This is available along with substantive data collected during adult individual interviews (including proxy interviews) in w_indresp.

4. DATA ACCESS

We request that researchers using the data notify us about errors, inconsistencies, and other problems with the data identified during their use of the data. Please send reports of errors and other problems to data@understandingsociety.org.uk. There is also a contact link on the online documentation pages. It would be helpful if you

would include a description of the problem, your log file, and information about how to contact you. We will communicate information to members of the Understanding Society users group or via Frequently Asked Questions on the *Understanding Society* web page about data: http://data.understandingsociety.org.uk/.

The data are released through the UK Data Archive (UKDA) in SPSS and Stata formats. While documentation is released through the UKDA, we encourage users to consult the *Understanding Society* webpage. The documentation will develop over time. We plan to be developing specific guides about major content areas such as the biomeasures or cognitive measures, and guides for issues that are frequently problematic for users such as selection of appropriate weights.

Most of the Wave 1 has been released according to the conditions of the regular UKDA End User Licence: https://www.esds.ac.uk/aandp/access/licence.asp. A version of the Wave 1 data has been released under conditions of the Special Licence (SL). Special Licence datasets are anonymised but contain more detailed information than End User Licence (EUL) data. The UKDA requires users to complete a set of forms with such detail as the intended use of the data. Researchers are asked to report publications resulting from the data.

Related *Understanding Society* releases are being prepared. One is a set of data products with information to link *Understanding Society* survey data with geographic units including Local Authority Districts, Area Classification for Output Areas, Travel to Work Areas, Westminster Parliamentary Constituencies, Rural-urban Indicators, Local Education Authorities, and Primary Care Trusts. For further information about these geographic units see (Office for National Statistics 2010). Many of these data releases are also done under Special Licence.

Users should acknowledge both the UKDA and the Institute for Social and Economic Research in any publications arising from analysis of the data. Notifications to ISER can be sent to info@understandingsociety.org.uk.

CITATIONS AND ACKNOWLEDGEMENTS

Readers wishing to cite this document should use these words.

McFall, Stephanie L. (2011). Understanding Society – The UK Household Longitudinal Study, Wave 1, 2009-2010, User Manual. Colchester: University of Essex .

People who participated in writing the documentation included Jon Burton, Peter Lynn, Olena Kaminska, Gundi Knies, Randy Banks, Cheti Nicoletti, Laura Fumagalli, Jakob Petersen, and Nick Buck. Many people participated in preparing and processing the questionnaires and data. From the information technology side we recognise the contributions of Paul Groves, Paul Siddall, Geoff Angel, Tom Butler, Jeannette Chin, Elaine Prentice-Lane, Muneeb Shaukat and Catherine Yuen. From the survey research team, we recognise Noah Uhrig, Sarah Budd, and Emily Kean. A small group was active in contributing code for derived variables and flagging issues in using the data. They include Jakob Petersen, Cara Booker, Alexandra Skew, Mark Bryan, Mark Taylor, and Alita Nandi.

5. REFERENCES

Berthoud, R., Fumagalli, L., Lynn, P., & Platt, L. (2009). *Design of the Understanding Society ethnic minority boost sample*, Understanding Society Working Paper No. 2009-02, Colchester: ISER, University of Essex. http://research.understandingsociety.org.uk/publications/workingpaper/2009-02.

Kenward, M., and J. Carpenter (2007), *Multiple imputation: current perspectives, Statistical Methods in Medical Research*, 16, 3, 199-218.

Lynn, P. (2009) Sample Design for *Understanding Society*. *Understanding Society Working Paper 2009-01*. Colchester: University of Essex. http://research.understandingsociety.org.uk/publications/working-paper/2009-01.pdf

Office for National Statistics (2010) *A Beginners' Guide to UK Geographies.* http://www.statistics.gov.uk/geography/beginners_guide.asp

Office of National Statistics (2010) 'Midyear population estimates 2009'. June 24, 2010 Edition, http://www.statistics.gov.uk/statbase/product.asp?vlnk=15106.

Ragunathan, E, T., Lepkowski, J. M., van Hoewyk, J., and Solemberger P. (2001), A Multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27, 1, pp.85-95.

Rubin, D. B. (1987): Multiple imputation for nonresponse in surveys. New York: Wiley.

Schafer, J. (1997): Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Taylor, M. F. (ed) (2010) British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: Universtiy of Essex.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681–694.